BIOINFORMATICS ORIGINAL PAPER
Vol. 30 no. 4 2014, pages 506–513
doi:10.1093/bioinformatics/btt704

Gene expression

Advance Access publication December 3, 2013

Joint estimation of isoform expression and isoform-specific read distribution using multisample RNA-Seq data

Chen Suo[1,†], Stefano Calza[1,2,†], Agus Salim[3] and Yudi Pawitan[1,*]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, [2]Department of Molecular and Translational Medicine, University of Brescia, Italy and [3]Department of Mathematics and Statistics, La Trobe University, Australia

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: RNA-sequencing technologies provide a powerful tool for expression analysis at gene and isoform level, but accurate estimation of isoform abundance is still a challenge. Standard assumption of uniform read intensity would yield biased estimates when the read intensity is in fact non-uniform. The problem is that, without strong assumptions, the read intensity pattern is not identifiable from data observed in a single sample.

Results: We develop a joint statistical model that accounts for non-uniform isoform-specific read distribution and gene isoform expression estimation. The main challenge is in dealing with the large number of isoform-specific read distributions, which potentially are as many as the number of splice variants in the genome. A statistical regularization via a smoothing penalty is imposed to control the estimation. Also, for identifiability reasons, the method uses information across samples from the same region. We develop a fast and robust computational procedure based on the iterated-weighted least-squares algorithm, and apply it to simulated data and two real RNA-Seq datasets with reverse transcription–polymerase chain reaction validation. Empirical tests show that our model performs better than existing methods in terms of increasing precision in isoform-level estimation.

Availability and implementation: We have implemented our method in an R package called Sequgio as a pipeline for fast processing of RNA-Seq data.

Contact: yudi.pawitan@ki.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 30, 2013; revised on November 6, 2013; accepted on November 28, 2013

1 INTRODUCTION

Alternative use of exons can form a number of combinations called splice variants, which can be used as templates for producing related but distinct proteins (Brett *et al.*, 2002). Alternative splicing has been observed among different tissue types in 90–95% of human genes (Matlin *et al.*, 2005; Wang *et al.*, 2008) and greatly diversifies the transcriptome. Many splice variants have been found to be implicated in a wide range of human diseases and functional roles (Nagao *et al.*, 2005; Wang *et al.*, 2003). For this reason, it is important to develop the technologies and statistical methods to distinguish and quantify different isoforms of the same gene.

To compute an absolute expression score—in reads per kilobase per million reads (RPKM) units—read counts are normalized against the transcript length and the total number of mappable reads (Mortazavi *et al.*, 2008). Typically, to estimate isoform abundance, read counts falling into a gene with multiple isoforms are modeled as a Poisson process with uniform sampling across each transcript (Jiang and Wong, 2009). But due to a number of factors, e.g. the 5′ or 3′ bias, local nucleotide composition effects—such as priming or GC bias—or other technical biases, read distribution might not be uniform (Howard and Heber, 2009). Indeed, empirical goodness-of-fit test for the Poisson model fitted under the uniform assumption shows that a majority of these models have poor fit. This would lead to bias in the isoform expression estimates.

Recent methods suggest estimating non-uniform read distribution from single-isoform genes (Howard and Heber, 2009; Li *et al.*, 2010; Wu *et al.*, 2010). These methods are rather limited, where either one distribution is used for all isoforms or a different distribution depending on length. In other words, all transcripts of the same length from all genes, regardless of what genes and whether the genes have single or multiple isoforms, are assumed to have the same read distribution. This ignores, for example, the local composition effect of the transcript. Furthermore, a recent study (Kozarewa *et al.*, 2009) did suggest that, as the method used for RNA library preparation introduces some amplification artifacts, the distribution of read coverage could be isoform-specific. In real data we do find that the distribution for different transcripts that share common length is not always the same. In addition, we observe that read distribution is highly correlated across samples. In Supplementary Report, Supplementary Figures S1 and S2 show typical examples of non-uniform read distributions with overabundance on the 3′ region. Interestingly, across samples, we find similarities in the shape of the read distributions even between different tissues.

There is evidence that the sample-to-sample similarity in non-uniform read distributions holds more generally across the genome; see Supplementary Figure S3 in the Supplementary Report.

The restriction imposed by the previous methods in estimating read distribution highlights one main difficulty: once we allow non-uniformity, in principle each transcript—even from the same

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

gene—to have its own specific pattern, would lead to a large number of distribution patterns, potentially as many as the number of transcripts. This makes the estimation highly non-trivial. To account for local composition effect, Hansen *et al.* (2010) adjusted for priming bias, where each read is reweighted based on its first few bases. If based on seven nucleotides (heptamer), they needed to calculate the proportion of reads starting with each specific heptamer, i.e. 4^7 frequencies. So reads beginning with a certain composition of seven nucleotides overrepresented in the heptamer distribution are down-weighted. Li *et al.* (2010) modeled the read counts depending on the specific composition of nucleotides along a gene, and developed a more complex Poisson linear and non-linear model to estimate the effect of a certain nucleotide occurring in the k^{th} nucleotides away from a given position j, by modeling the count of reads starting at position j. The gene-expression levels and the coefficients of the effect of surrounding nucleotide are optimized iteratively.

Cufflinks (Trapnell *et al.*, 2010), one of the mostly commonly tools used to deal with sequence-specific biases problem in isoform expression estimation, assumes uniform read distribution in its basic model, but provides an *ad hoc* correction of bias step (Roberts *et al.*, 2011). It also estimates positional bias, which measures whether fragments are preferentially located toward either ends of the transcripts. Unlike the base-level bias correction method, our idea is to model the isoform-specific read distribution and expression jointly. The observed sample-to-sample similarity suggests that combining data from different samples makes read distribution identifiable. Our approach automatically allows for the local composition effect without any need for explicit modeling. Such a joint model is more natural, so we expect it to lead to better performance. The algorithm of iteratively estimating isoform expression and read distribution has some similarity to the idea in a recent published program NURD (Ma and Zhang, 2013), where a global bias curve for all genes and a local bias curve for each gene is estimated using non-parametric models. In this study, we compare the estimation accuracy of our model against Cufflinks, NURD and the method under uniform read distribution assumption, using a set of simulated datasets, and also apply our proposed method to two real datasets.

In summary, the purpose of this article is to describe a method for joint estimation of isoform-level expression and isoform-specific read distribution. Allowing for isoform specificity automatically deals with positional bias and local composition effects, but it is challenging, as we then have to estimate as many distributions as the number of isoforms in the genome. Regularization via a smoothing penalty is used to control the estimation of the read distributions. Primary results show that, our approach provides substantial improvement on the quality of model fitting and improves the sensitivity in isoform-level differential expression analysis, compared with the method based on uniform assumption, Cufflinks and NURD.

2 METHODS

In previous methods (e.g. Jiang and Wong, 2009), a model is fitted gene by gene separately. Instead of genes, we consider a more natural model based on non-overlapping 'transcriptional units', each of which is defined as a union of all overlapping transcripts. A transcriptional unit may possibly contain several overlapping genes. For example, >10 distinct genes lie between position 88 022 280 and 88 277 580 on mouse chromosome 1, including a UDP glucuronosyltransferase 1 family; this family comprises eight transcripts that are annotated with different gene names in RefSeq. If a read is mapped to a region of overlapping genes, it is not possible to decide which gene the read comes from. If the genes are treated separately, the reads would be doubly counted, resulting in falsely higher expression level.

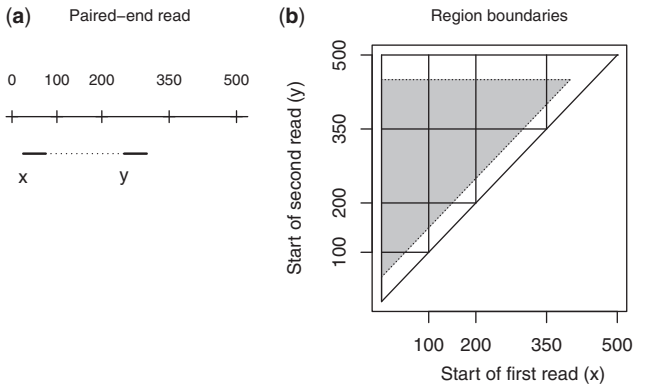

Fig. 1. (**a**) A schematic illustration of a paired-end read that aligns to a transcript. The values (100, 200, 350, 500) are chosen as an example of region boundaries; the read length is 50. The 'x' and 'y' mark the aligned starting positions of the reads. (**b**) A 2D representation of possible aligned positions of paired-end reads in a transcript, where each read pair is represented by a point in the shaded area. Width of the strip at the bottom and the top of the triangle equals to the read length. The number of pairs that fall in the regions within the shaded area is recorded as the read-count data

2.1 Read-count data and the general model

To facilitate fast computations, we first summarize the number of reads that align to distinct subregions of a transcriptional unit. Whenever we say regions or subregions, we refer to exons and junctions. Let y_{ri} be the number of fragments from individual i that fall in region r, i.e. the corresponding aligned reads in region r. This counting procedure is obvious for single-end reads, as we can simply count the number of reads that align in each region. For paired-end reads, we construct a two-way table such that the $(i, j)^{th}$ entry records the number of fragments whose first reads fall in the i^{th} region and second reads in the j^{th} region. Hence, in the pair-end case, a 'region' r is naturally defined by a pair of subregions (i, j). With this understanding, the same notation y_{ri} applies for both single- and paired-end reads. Hereafter, unless needed for clarity, we simply use the term 'region' for both single- and paired-end data. For a transcriptional unit g with J known isoforms, let θ_{ji} be the expression level of isoform j in individual i. The main statistical problem is to estimate transcript abundances θ_{ji}'s from the read-count data y_{ri}'s.

Because of its complexity, we describe the paired-end case in detail; as explained later in the text, the single-end case is a special case of the paired-end. Each paired-end read that aligns inside a transcript is characterized by the starting point of each read; see Figure 1a. It is then convenient to represent all read pairs that align inside the transcript by the shaded area of Figure 1b. For a given transcript, the implied fragment length associated with a read pair at (x,y) is equal to $\ell_j(x, y) = (y - x + k)$, where k is the read length. Because the fragment length has a certain distribution, e.g. it is normally distributed with mean 300 bases, it affects what x and y values are possible. We have also noted

previously the importance of assuming non-uniform read distribution across the transcript. Thus, the distribution of read pairs over a transcriptional unit can be represented by a 2D point process whose intensity is determined by the total number of mapped reads, transcript abundances, local non-uniformity effects and fragment-length density. Specifically, by adding up the contribution of multiple isoforms, we model the point-process intensity at (x,y) as

$$h_i(x,y) = w_i \sum_{j=1}^{J} \theta_{ji} c_j(x,y) f(\ell_j(x,y)),$$

where w_i is conventionally the total number of mapped reads divided by 10^9, $c_j(x,y)$ is the transcript-specific non-uniformity effect, $\ell_j(x,y)$ is the fragment length implied by the (x,y) position in transcript j and $f(\cdot)$ is the fragment-length density. Now, for each region r, let R_r be the corresponding area defined by the region boundaries in the shaded area of Figure 1; to be clear, from the figure, the shape of R_r could be a triangle, a rectangle or a rectangle-minus-triangle. The expected number of read pairs in R_r is

$$\begin{aligned}\lambda_{ri} &\equiv \int\int_{(x,y)\in R_r} h_i(x,y)dxdy \\ &= w_i \sum_{j=1}^{J} \theta_{ji} \int\int_{(x,y)\in R_r} c_j(x,y) f(\ell_j(x,y))dxdy .\end{aligned}$$

In practice, $f(\cdot)$ is either assumed known, e.g. normal with certain mean and variance, or estimated using only read pairs from single-isoform transcriptional units. We shall estimate $c_j(x,y)$ jointly with transcript abundances θ_{ji}'s. To get some simplifications, we assume that $c_j(x,y) \approx c_{rj}$ for $(x,y) \in R_r$, so the integration is always on a known function, and the model can be written as

$$\lambda_{ri} = w_i \sum_{j=1}^{J} \theta_{ji} c_{rj} \int\int_{(x,y)\in R_r} f(\ell_j(x,y))dxdy , \quad (1)$$

$$\equiv w_i \sum_{j=1}^{J} \theta_{ji} c_{rj} L_j x_{rj} , \quad (2)$$

where we define

$$\begin{aligned}L_j &\equiv \sum_r \int\int_{(x,y)\in R_r} f(\ell_j(x,y))dxdy \\ x_{rj} &\equiv \frac{1}{L_j} \int\int_{(x,y)\in R_r} f(\ell_j(x,y))dxdy .\end{aligned}$$

To interpret these quantities, we can see that if the fragment length is fixed, then L_j is the total length of transcript j minus the fragment length, so in general L_j is the effective length of transcript j. Furthermore, x_{rj} can be interpreted as the proportion of read pairs in region r under uniform read distribution. By definition $\sum x_{rj} = 1$ for every j. Given a transcript annotation database, the full collection of L_j's and x_{rj}'s need to be evaluated only once.

For the single-end case, the region r is an interval coinciding with exons or junctions. In this case, only the first reads are counted, so the double integrals in (1) reduce to the first integral over x, and x_{rj} is now the ratio of the length of region r relative to the total, and the same general model (2) applies.

2.2 Uniform read distribution

Consider firstly the model under uniform read distribution assumption, so from (2), with $c_{rj} \equiv 1$, we have

$$\lambda_{ri} = w_i \sum_{j=1}^{J} L_j x_{rj} \theta_{ji} \quad (3)$$

It is typically assumed also that y_{ri} has Poisson distribution with mean λ_{ri}. This simple model (3) has been developed and used in several previous RNA-Seq studies (Jiang and Wong, 2009; Mortazavi *et al.*, 2008) and will be referred to as the standard method.

2.3 Non-uniform read distribution

For a specific transcriptional unit, we find that read distributions across samples are similar, as we have seen in Supplementary Figure S1. Dividing length of a transcriptional unit into bin width of 200 bp, the read counts in each bin across samples have similar patterns, as shown in Supplementary Figure S2 in the Supplementary Report. If expression values are estimated from a single sample, it would not be feasible to discover and unveil the underlying true read distribution. However, by combining reads from multiple samples, it is possible to estimate the read distribution, as multiple observed data points are available to determine the read intensity of a region.

In this general case, we also start with the assumption that y_{ri} is Poisson with mean λ_{ri} in (2). Joint estimation of θ_{ji}'s and c_{rj}'s will be done using the maximum likelihood approach. The likelihood function is given in Section G of the Supplementary Report. First we can see that, given isoform-specific read intensity c_{rj}'s, the isoform expression θ_{ji}'s can be estimated from a linear model:

$$\lambda_{ri} = w_i \sum_j (c_{rj} L_j x_{rj}) \theta_{ji} \equiv \sum_j a_{rj} \theta_{ji} , \quad (4)$$

where $a_{rj} \equiv w_i c_{rj} L_j x_{rj}$. For identifiability, we set $\sum_r (c_{rj} x_{rj}) \equiv \sum x_{rj} = 1$ for every j. With this restriction, it is clear that the assumption of uniform read distribution implies $c_{rj} = 1$ for all r and j, and the general model (4) reduces to the standard model (3). Thus, the joint estimation can be performed iteratively as follows:

A. Given c_{rj}, estimate θ_{ji} sample by sample using model (4).
B. Given θ_{ji}, estimate c_{rj} using this model:

$$\lambda_{ri} = w_i \sum_j (L_j x_{rj} \theta_{ji}) c_{rj} . \quad (5)$$

C. Iterate steps A–B until convergence.

The iterative scheme can be recognized as a block Gauss–Seidel method. The iteration looks simple, but in practice we have to use various estimation techniques to ensure a robust and fast computational procedure. In step A, given c_{rj} to estimate λ_{ri}, we use a generalized linear model with an identity link function. To make the estimation robust to outliers, we perform iterative-weighted least-squares with robust modification to deal with potential outliers (Pawitan, 2001; Chapter 6.7). The explicit steps and the computation of variance of the estimates are given in Section G of the Supplementary Report.

The problem of estimating c_{rj} in step B is more complex than the estimation of θ_{ji} in step A, as now we are dealing with many more parameters. It is possible to estimate read intensity c_{rj} for each region separately. However, intuitively we do not expect read intensity to change dramatically between adjacent regions. So, to allow the possibility of smooth transition between neighboring regions, we consider a model with smoothness penalty. This is done using a generalized linear mixed model with isoform-specific read intensity as correlated random effects (Pawitan, 2001; Chapter 18).

Overall, the likelihood estimation with smoothness penalty is equivalent to a constrained optimization problem. The iteration scheme follows a block Gauss–Seidel method; its convergence is guaranteed in this case because both the likelihood and the penalty constraint are convex (Grippo and Sciandrone, 2000). The convexity of the log-likelihood function has been shown by Wu *et al.* (2010) and Jiang and Wong (2009).

The penalty function is based on the Gaussian distribution, so it is convex. Indeed, we checked that the Hessian matrix of the negative log-likelihood function is numerically positive definite for every model fitted. With the real mouse data, computations for 97% of the transcriptional units converge within 10 iterations; the rest have slower convergence, most likely due to low coverage.

2.4 Availability and implementation

The algorithm described is implemented in the R statistical programming language (http://www.r-project.org/). The package, called Sequgio, is available in Bioconductor's library, with a corresponding vignette. It could be part of an R-based pipeline for measuring differential expression of isoforms using RNA-Seq datasets. The package allows users to load in reads mapped by any alignment program, such as Burrows-Wheeler Alignment tool (BWA) (Li and Durbin, 2009), Tophat (Trapnell *et al.*, 2009) or Bowtie (Langmead *et al.*, 2009). It is freely available on the web at http://www.meb.ki.se/~yudpaw.

3 DATASET AND PREPROCESSING PROCEDURES

3.1 RNA sequencing data—mouse tissues and human brain tissue

We use RNA sequencing samples published in Mortazavi *et al.* (2008) (NCBI Short Read Archive SRA001030). The samples consist of three mouse tissues: brain, liver and skeletal muscle. Two replicates of the same tissue-type are sequenced on the Illumina-Solexa platform. For each sample, there are ~20–30 million reads. Among all the transcripts defined in UCSC database, 36.0% are from multiple isoform-genes. The majority of the 3939 multi-isoform genes in mouse have between two and four isoforms. Specifically, 2629 genes have 2 isoforms, 773 have 3 isoforms, 273 have 4 isoforms, 264 have >4 isoforms and the maximum number of isoforms is 15.

A second RNA-Seq dataset with read length of 50 nt are obtained from the Microarray Quality Control project (MAQC), Gene Expression Omnibus accession GSE19166 (MAQC consortium *et al.*, 2006). Three technical replicates from Human Brain Reference of the same sample are sequenced by Illumina. Each replicate contains ~7–8 million reads. A benchmark dataset consisting of 1044 transcripts analyzed by TaqMan quantitative reverse transcriptase–polymerase chain reaction (qRT–PCR) is used for evaluating the performance of expression quantification methods. However, bear in mind that the TaqMan assay is not a perfect validation tool. First, although the four replicates of qRT-PCR data are generated from the same batch of Human Brain Reference RNA, those four replicates and the three replicates for RNA-Seq data are from different aliquots of the same RNA sample. Comparisons will thus be based on the average expression estimate of the qRT-PCR replicates and the average expression estimate of RNA-Seq replicates. Second, it is difficult to discriminate the expression of distinct isoforms using qRT-PCR. An annotation with one-to-one match between genes and isoforms in the PCR experiment is available in Supplementary Table S2 of the MAQC project (MAQC consortium *et al.*, 2006). But we find that it might not be reliable, as although the annotation may relate a gene to one of its isoforms, we can tell from the RNA-Seq and qRT-PCR estimates that it might be the other isoform or both expressed. So we decide to use the commonly used UCSC annotation NCBI36/hg18 and summarize isoform-level expression to gene-level before comparing the estimates from the different platforms.

3.2 Simulation procedure

To study the performance of the proposed method, we conduct simulations to compare results of Sequgio with (i) the transcriptional unit- and (ii) the gene-based standard methods, (iii) Cufflinks (version 2.0.2) with bias correction (Trapnell *et al.*, 2010) and (iv) NURD (Ma and Zhang, 2013). The basic model in Cufflinks assumes the uniform read distribution, but for our comparisons we have used their suggested bias-correction step to account for non-uniform read distributions. The simulations are based on all transcripts in the reference annotation, and single-end reads are simulated.

We first consider a model-based simulator. This is useful as evidence that all the methods have been implemented properly. To be realistic, the parameters are based on real data and performed as follows: first, the expression value and read distribution for each of the isoforms are estimated from the RNA-Seq data in Mortazavi *et al.* (2008), along with known total number of reads and length of exons and junctions. The expected number of reads for every region is then calculated using (4). Then, counts along the genomic regions in 10 samples are drawn from the Poisson distribution with the expected read counts as the mean. The output of each run of the simulation is read-count data in exons and junctions. We then convert these counts to a BAM/SAM file format, which is the input format accepted by Cufflinks and NURD. Read starting points would be randomly assigned within a region, i.e an exon or a junction, as long as the sum of reads equal to the simulated count for the region. For every randomly generated read, a corresponding cigar is generated and the genome annotation is queried accordingly to collect the corresponding sequence. We then format the reads to SAM style according to the assigned mapped positions. To check the effect of the alignment or mapping step, we further convert the BAM file to an unaligned FASTA file format and reprocess the BAM file with an aligner. Converting read counts into BAM and FASTA is performed using Python and the pysam (https://code.google.com/p/pysam/) and Bio modules.

We then use a second simulator called RNASeq ReadSimulator that is fully independent of our model (http://www.cs.ucr.edu/~liw/rnaseqreadsimulator.html). This simulator program generates sequencing reads according to certain parameters. Crucially, it allows the users to specify the positional bias at isoform-level such that read distribution would be non-uniform. We use the observed read distributions and expression levels in Mortazavi *et al.* (2008) to capture the realistic patterns of positional bias and expression distribution. The expression level of each transcript (in RPKM units) is calculated by normalizing against the total number of reads within a sample and the effective transcript length. The raw simulated data are unmapped reads in FASTA format, which are then processed through the pipelines for the different methods.

Among the methods with sequence bias correction, NURD considers a gene as one unit, whereas the others consider overlapping genes together. So, to investigate the necessity of modeling overlapping genes simultaneously, we compare the

performance of the methods on transcriptional units that contain multiple genes.

As a measure of closeness, resulting isoform-level expression estimates (O) are compared with the predetermined true expression values (E) using an absolute proportion error

$$e = |O - E|/E.$$

This measure of distance gives a clear idea how close the estimates are to the true values. We then calculate the median value across samples for a transcriptional unit. To summarize the comparative procedures, we compare Sequgio, Cufflinks, NURD and the standard method using simulated reads from four sets of simulation: (A) mapped reads (turned into BAM and SAM file for Cufflinks and NURD, respectively), (B) unmapped reads (FASTA) from model-based simulator, (C) unmapped reads generated by RNASeqReadSimulator where expression levels are simulated based on real data and (D) a subset of (C) only contain transcriptional units formed by more than one gene.

4 RESULTS

4.1 Simulation

For mapped reads from model-based simulator (scenario A), we find Sequgio, Cufflinks and NURD has an overall correlation coefficient with true values of 0.96, 0.93 and 0.90, respectively. For single-isoform transcriptional units, correlation is all 0.99 for the three algorithms. For multi-isoform transcriptional units, Sequgio estimates have a correlation of 0.90 with the true values, compared with 0.85 and 0.78 for Cufflinks and NURD estimates. These high correlation values are obtained across all types of the simulated data, indicating the estimation procedures work as expected.

The advantage of the joint model is expected to be more obvious when read distribution is deviated severely from uniformity. We show this in a simulation study given in Section B of the Supplementary Report, from which we note that there is no loss of performance of Sequgio in terms of model fitting when the read distribution is close to uniform. So, we stratify the transcriptional units to those that moderately and severely deviate from uniform according to whether a non-uniformity deviance measure is less or greater than its median value. The deviance measure is defined as the averaged squared difference between true read intensity and uniform intensity.

The overall simulation results are summarized in Table 1. Among units whose non-uniform deviance is less than the median value, the median proportion errors are 4.0, 12.5, 14.1, 5.0 and 6.9% for Sequgio, the transcriptional unit- and gene-based standard method, Cufflinks and NURD, respectively, where Sequgio has the lowest error. For units whose non-uniform deviance is larger than the median, the median proportion errors are 4.6, 5.8, 7.0, 5.5 and 6.6%, and Sequgio estimates are again the closest to the true values. Differences between methods are larger in transcripts with the severe non-uniform read distribution, especially when Sequgio is compared against the standard methods. We observe in particular that the gene-based standard method performs worst among all the methods. This is in line with what we expect, as the gene-based method cannot distinguish reads falling into overlapping genes entirely. In the following simulations, we will not include the gene-based standard method.

Table 1. Results of comparing Sequgio, Cufflinks, NURD, the transcriptional-unit and gene-based standard method from four simulation settings

Number of transcriptional units (N)	Median proportion error	
	Moderate	Severe
(A) Model-based simulator (BAM)	4082	4081
Sequgio	4.6%	4.0%
Standard	5.8%	12.5%
Gene-based standard	7.0%	14.1%
Cufflinks	5.5%	5.0%
NURD	6.6%	6.9%
(B) Model-based simulator (FASTA)	4082	4081
Sequgio	6.1%	5.7%
Standard	13.5%	14.1%
Cufflinks	27.7%	31.2%
NURD	8.4%	7.8%
(C) RNASeqReadSimulator (FASTA)	4877	4876
Sequgio	5.9%	5.2%
Standard	7.1%	10.5%
Cufflinks	6.7%	6.2%
NURD	6.3%	5.9%
(D) RNASeqReadSimulator (FASTA) (multigene transcriptional units)	386	159
Sequgio	8.3%	9.2%
Standard	8.8%	10.9%
Cufflinks	12.1%	14.1%
NURD	15.9%	19.0%

Note: Unless otherwise specified, the standard method is fitted for a transcriptional unit, not gene-based. Within each transcriptional unit, the median proportion error of expression level for every transcript across all samples is computed. We stratify transcriptional unit by its read distribution moderately and severely deviated from uniform.

To analyze simulated data in FASTA format (scenario B), we first have to map reads to the reference genome by an alignment program, such as Tophat, before doing any expression estimation. This setting tells us how well the methods deal with various alignment issues; for example, we find that reads simulated on the negative strand may end up mapped to the positive strand. In addition, filtering procedures of mapped reads, such as the threshold of mapping quality score, may be applied differently by expression quantification tools. Compared with the results from reads simulated at the level of unmapped sequences, performance of all estimators is affected to some extent by the mapping procedure. Cufflinks is the most poorly affected, whereas Sequgio with joint modeling has a consistently good performance among the estimators.

Processing data simulated from RNASeqReadSimulator (scenario C) can be considered as a proper test for Sequgio, as the simulator is from an independent source and its output are unmapped reads, so abundance quantification may be affected by mapping and filtering procedure. Sequgio still maintains high-quality performance with the median proportion error <6%.

When the comparison is restricted to a subset of transcriptional units that contain multiple genes (scenario D), NURD does not function as well as in the other scenarios. NURD treats these genes independently, resulting possible double counting, arbitrary assignment or removal of reads. Thus, overall Sequgio performs better than the other methods.

4.2 Sensitivity in differential expression analysis

To investigate how much power one can gain in estimating read distribution by integrating multiple samples, we perform a simulation study for detecting differential expression levels estimated by the standard and the Sequgio method. Our model simulates the number of reads given exon length and read distribution from a real example, gene Cinp that contains six exons of lengths between 100 and 1500 bp. Two isoforms NM_026048 and NM_027223, one with 5 exons and the other with 6. We take data of this gene because its P-value of the goodness-of-fit test for models fitted using the standard method is around the third quartile (see Section 4.3). Also it is an illustration of a gene where read distribution differs between the isoforms, one transcript with a uniform read distribution and the other not. The number of samples in each tissue group is set to be six and the expression fold change δ of two tissue types varies from 1 to 1.3. The simulation is performed following the same procedures as described for model-based simulator, but for only one gene. The expression levels between tissues are then compared using *t*-test with unequal variances. For each value of δ, we generate 100 simulation sets and estimate the power to detect differentially-expressed (DE) transcripts. Results are shown in Figure 2. In panel (a), the power analysis shows that using Sequgio we are able to identify more true DE transcripts. The gain in power can be as much as 20% at fold-change $\delta \sim 1.2$. Panels (b) and (c) indicate that the read distributions are estimated well.

4.3 Real RNA-Seq datasets

To test the performance of our method on real data, we analyze the publicly available mouse tissue dataset (Mortazavi *et al.*, 2008). For each gene whose expression is modeled by the standard method, we compute a goodness-of-fit statistic across regions to test the hypothesis that the assumption of uniform read distribution is adequate. The *P*-values from the goodness-of-fit test are used to access the model fitting. After correcting for multiple testing, 68.5% of the models have a *P*-value < 0.05, indicating that a majority is poorly fitted using the standard method. Overall, χ^2 statistics are reduced for 70.3% of the models, indicating that Sequgio improves model fitting substantially. In Section D of the Supplementary Report, we show the reduction in χ^2 statistics in detail for a gene with a median *P*-value from the goodness-of-fit test. We also calculate the average absolute deviation from uniform read intensity within every transitional unit in the mouse tissue data. The median value across units is 0.214, which can be interpreted as that the average estimated read intensity deviates 21.4% from uniformity.

We next report the differential expression analysis of the tissues, and compare the gene-level versus the isoform-level differential expression. Overall the gene expression is equal to the sum of isoform abundance. Differential expression is defined using fold changes (FCs) and false discovery rate from moderated

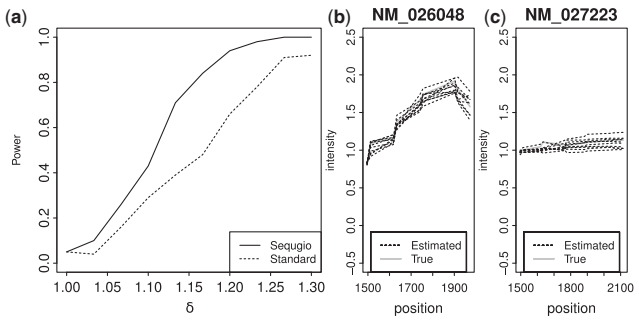

Fig. 2. Figure on the left, (**a**) power as a function of δ, the mean difference between groups. (**b** and **c**) True and estimated read distribution for two isoforms

Welch *t*-test (Demissie *et al.*, 2008). Specifically, DE genes are those with FC > 3, either over- or underexpressed, and false discovery rate < 0.01. Expression values are analyzed on $\log(x + 1)$ scale because of the zeros in the data. Transcripts with variance < 0.001 are removed. Only genes with multiple isoforms are considered. Table 2 summarizes results of the differential expression analysis. Comparing brain and liver tissues, 18.7% of the 30 140 tested transcripts shows differential expression. Among genes with DE isoforms, 20.4% does not show gene-level differential expression pattern, indicating that testing on isoform-level is necessary. Overall, 246 transcripts show strong differential expression between brain and liver tissues with an FC > 10 and an average expression value >100 in at least one of the tissues.

To further examine the genes identified as DE between tissue types, we look at a known tissue-specific gene Mecp2. Mutations in Mecp2 are the essential cause of the Rett syndrome, a neurodevelopmental disorder of the brain (Amir *et al.*, 1999). It is involved with brain development and neuron function. In one of Mecp2's isoforms ENSMUST00000123362, the estimated average expressions using Sequgio for the brain, liver and muscle tissues are 5.78(0.0003), 0.63(0.0014) and 0.00(0.008), respectively, with standard error indicated in the parenthesis. Using the standard method, the estimated expressions are 3.96(0.002), 0.94(0.001) and 1.69(0.006), respectively. The standard error of expression levels is small. So, the difference between the expression is larger in Sequgio's estimates.

4.4 Validation with mouse RT-PCR data

We compare results derived by our method with RT-PCR quantification as in Zheng and Chen (2009), where RT-PCR is performed to assay transcripts' relative expression levels in the three mouse tissues. For each transcript, a relative expression ratio is computed between brain and muscle, and between brain and liver. For all seven genes annotated with the Alternative Splicing and Transcript Diversity (ASTD, http://www.ebi.ac.uk/astd/) and Ensembl (http://www.ensembl.org/index.html) databases, four transcripts can be matched or partially matched to a gene in the UCSC annotation that we are using. Isoform expression values are computed using our method in the three mouse tissue RNA-Seq data obtained from Mortazavi *et al.* (2008). We find that in all comparisons except for transcript

Table 2. Results of differential expression analysis between tissues

Comparison (A) versus (B)	Br versus L	Br versus M	L versus M
Transcripts DE (1)	18.7%	15.1%	5.4%
Non-DE at gene level among (1)	20.4%	13.8%	12.0%
Transcripts FC > 10 and highly DE	246	187	201
Transcripts upregulated in (A)	127	122	126

Note: The first row lists the percentage of DE isoforms. But genes with DE isoforms can be non-DE at gene level. Percentage is summarized in the second row. Last two rows show the number of transcripts with both FC > 10 and highly expressed; the number of transcripts upregulated in tissue (A). Br, L and M stands for brain, liver and muscle.

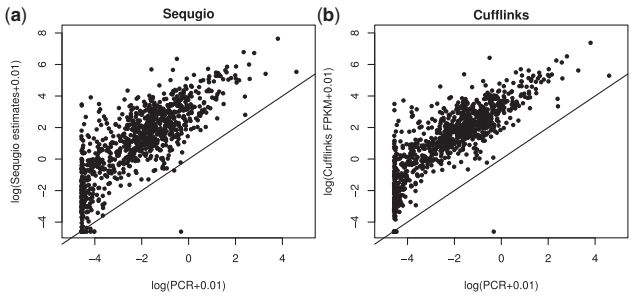

Fig. 3. (**a**) Estimates from Sequgio are plotted against the expression levels obtained by RT-PCR. (**b**) Estimates from Cufflinks are plotted against the expression levels obtained by RT-PCR. Expressions values are on the log scale

ENSMUST00000057185 (Pcdh1) in brain/liver and transcript ENSMUST00000115609 (Comt), they show the same direction of distinct expression patterns (Supplementary Table S2, Supplementary Report).

Not all of our results are in agreement with RT-PCR, although the discrepancies could be due to errors in annotation database and variability in isoform structures. For example, one of the amplified transcript ENSMUST00000115599 (Pcdh1) is no longer in the Ensembl database and has not been mapped to any new identifier. Our result of the expression estimation, although annotated to the other transcript of the same gene, is closely matched with the expression of ENSMUST00000115599. Results produced from our method may be improved with a more consolidated and accurate annotation database. In addition, we note that the relative expression ratio of ENSMUST00000115609 is consistent with the result delivered by Zheng and Chen (2009) based on the sequencing data, i.e. differential expression pattern is barely detectable in brain/liver and it is positive direction in brain/muscle.

4.5 Validation with RT-PCR data in human brain tissue

We begin by examining the expression estimates from Sequgio; see Figure 3. The correlation of the average expression estimate and the average results of qRT-PCR analysis was 0.82. This is largely in line with the result from Cufflinks, which shows a correlation of 0.84 after bias correction. As we have pointed out in the simulation study, correlation may not be a good statistic for comparing the performance of the two algorithms. The χ^2 statistic would be better, but, as the expression values from the qRT-PCR experiment are not the true absolute expression, we are not able to calculate the statistic here. Indeed, we can see a systematic deviation from the line of identity in the scatter plot in Figure 3. Also, we examine the expression on the log scale, instead of on the original scale as shown by Roberts *et al.* (2011) in their supplementary report. This is because the correlation coefficient can be severely affected by outliers, which are common in the original expression estimates; see Supplementary Figure 6S in the Supplementary Report. The difference on scale explains why the correlation we obtain is different from in previous studies (Glaus *et al.*, 2012; Roberts *et al.*, 2011).

5 DISCUSSION

In this article, we introduce a novel method using RNA-Seq data from multiple samples to estimate the isoforms expression, taking into account non-uniform read distribution. Through simulations that model read-count data from non-uniform distribution, we demonstrate that our method improves accuracy in the expression quantification. When read distribution deviates dramatically from uniform, there is a striking improvement in accuracy of the expression estimation. Furthermore, our method can be easily adapted for use with any next-generation sequencing technology and mapping program.

Sequgio uses a fundamentally distinct method to estimate read distribution. First of all, it does not assume any relationship between sequencing bias and relative position of a certain nucleotide in a fragment. It uses an explicit and transparent model with isoform-specific read intensity to simultaneously correct different sources of bias. Second, Sequgio is able to estimate isoform-specific read distribution as long as count-level data are available for a single transcriptional unit. In contrast, to produce nucleotide-specific bias weights, Cufflinks requires the nucleotide-level information of all genes. The same global nucleotide distribution is then used for all fragments. But estimation of nucleotide bias weights might be sensitive to which and how many single-isoform genes are taken into calculation.

The main assumption of the joint model is that non-uniform read distributions can be identified using information across samples, given that the read distribution is consistent across samples. In section A of the Supplementary Report, we have presented evidence that the sample-to-sample similarity holds generally in the genome even between different tissues. In practice, we recommend users to follow the procedure to check the consistency especially when pooling information from two biological groups, e.g. diseased versus healthy. If not consistent, the estimation should be done separately. In the application to human brain tissue data, although only three samples are available, we see Sequgio performs fairly well based on the correlation with Cufflinks' and RT-PCR estimates. When there are <10 samples, we would recommend using them all in estimation. On the other hand, if a large number of samples are available and the computational system is limited, it would be useful to consider a two-staged procedure: (i) in the first stage the read

intensities are estimated from a subsample, and (ii) in the second stage these intensities are fixed, so only expression levels need to be estimated.

Gold standards for transcript-level expression are difficult to obtain experimentally. Improvement by our method is shown mostly by empirical means via the goodness-of-fit χ^2 statistics. We also rely on simulations and limited isoform-level RT-PCR data to assess the accuracy of our results. In the simulations, we consider both a non-uniform distribution and a slight deviation from uniformity, and all parameter values are those we estimate from the real data, so they are a fair testing procedure.

One limitation of many current methods including our own is that all isoforms for a gene are assumed to be known. Because of the huge amount of information from different isoform-level annotation database and complex structure of transcriptome, the current annotation is incomplete. We suspect this might partially cause discrepancies in the RT-PCR validation. Most biological annotation databases may be updated almost every week. Some databases will be closed and merged with others, e.g ASTD are integrated in Ensembl database. There is a need to develop a reliable and comprehensive mega annotation database. But it is worth emphasizing that with the RNA-Seq mouse tissue data used in this article, mapped counts and ~18 000 genes identified are comparable with results from those studies that use the same dataset (Jiang and Wong, 2009; Mortazavi *et al*., 2008).

Funding: Swedish Research Council (in part) 621-2009-5900.

Conflict of Interest: none declared.

REFERENCES

Amir,R.E. *et al.* (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.*, **23**, 185–188.

Brett,D. *et al.* (2002) Alternative splicing and genome complexity. *Nat. Genet.*, **30**, 29–30.

Demissie,M. *et al.* (2008) Unequal group variances in microarray data analyses. *Bioinformatics*, **9**, 1168–1174.

Glaus,P. *et al.* (2012) Identifying differentially expressed transcripts from RNA-Seq data with biological variation. *Bioinformatics*, **28**, 1721–1728.

Grippo,L. and Sciandrone,M. (2000) On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Oper. Res. Lett.*, **26**, 127–136.

Hansen,K.D. *et al.* (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.*, **38**, e131.

Howard,B.E. and Heber,S. (2009) Towards reliable isoform quantification using RNA-SEQ data. In: *Int. Conf. on Bioinformatics and Biomed.*

Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.

Kozarewa,I. *et al.* (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods*, **6**, 291–295.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

Li,J. *et al.* (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol.*, **11**, R50.

Ma,X and Zhang,X. (2013) NURD: an implementation of a new method to estimate isoform expression from non-uniform RNA-Seq data. *BMC Bioinformatics*, **14**, 220.

Matlin,A.J. *et al.* (2005) Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.*, **6**, 386–398.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Nagao,K. *et al.* (2005) Detecting tissue-specific alternative splicing and disease-associated aberrant splicing of the PTCH gene with exon junction microarrays. *Hum. Mol. Genet.*, **14**, 3379–3388.

Pawitan,Y. (2001) *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, the United States.

Roberts,A. *et al.* (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.*, **12**, r22.

MAQC Consortium. *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.

Trapnell,C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Wang,H. *et al.* (2003) Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics*, **19**, 315–322.

Wu,Z. *et al.* (2010) Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. *Bioinformatics*, **27**, 502–508.

Zheng,S. and Chen,L. (2009) A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. *Nucleic Acids Res.*, **37**, e75.